基于模糊 SVM 模型的入侵检测分类算法

汪 生[1]，金志刚[2]

(1. 中国北方电子设备研究所，北京 100191；2. 天津大学 电气自动化与信息工程学院，天津 300072)

摘 要：为解决入侵检测分类遇到的训练样本数量少、分类准确率低的问题，提出基于模糊支持向量机的多级分类机制。该分类机制首先训练模糊 SVM 模型将数据粗分为正常与攻击大类，然后采用 DBSCAN 算法产生细分模型进行攻击子集的自动聚类，将有关数据细分得到攻击的具体细类。在机制设计中，优化了隶属度函数的计算、设计了数据标准化与归一化等过程，并训练了高效分类器。实验表明，针对网络入侵检测数据中常见的孤立点干扰、噪声多、并且负样本占比多的网络业务数据集，新算法在保持分类准确率高的前提下，分类过程需要的计算时间较短。

关键词：模糊；SVM；入侵检测；分类

中图分类号：TP393 doi: 10.19734/j.issn.1001-3695.2018.08.0565

IDS classification algorithm based on fuzzy SVM models

Wang Sheng[1], Jin Zhigang[2]

(1. *Northern Institute of Electronic Equipment of China*, *Beijing* 1000191, *China*; 2. *School of Electronic & Information Engineering*, *Tianjin University*, *Tianjin* 300072, *China*)

Abstract: In order to solve the problem of small number of training samples and low classification accuracy in intrusion detection classification, this paper proposed a multi-level classification mechanism based on fuzzy support vector machines (FSVM) . This classification mechanism firstly trained the fuzzy SVM model to divide the data roughly into normal and attack categories, and then used DBSCAN algorithm to generate subdivision model for automatic clustering of attack subsets, and then subdivided the relevant data into specific classes of attack. In the mechanism design, the calculation of membership function has been optimized, the process of data standardization and normalization has been designed, and the efficient classifier has been trained. Experiments show that the new algorithm requires less computing time in the process of classification under the premise of high classification accuracy, aiming at the network service data sets with frequent isolated point interference, much noise and a large proportion of negative samples in the network intrusion detection datasets.

Key words: fuzzy; SVM; IDS; classification

0 引言

日益增长的网络用户数量导致网络安全问题更为突出。根据 CNNIC 第 41 次报告：截至 2017 年底，我国网民数量接近 8 亿，并且大量用户使用网络金融等高价值应用[1]。面对网络安全攻击带来的数据泄露、网络诈骗、拒绝服务等问题，被动防御和主动探测等网络安全技术成为研究热点。

入侵检测是主动网络安全的重要措施，也是安全防御的关键手段，一般通过对数据包进行分析和鉴别，以采取合适的防御联动措施。由于网络行为日益复杂，需要灵活高效的识别与分类算法对数据进行分析。

面向网络入侵检测的状态估计与分类可以采用支持向量机(supporting vector machine, SVM)模型[2]。文献[3]提出主成分分析（principal component analysis, PCA）与 SVM 机制联合进行入侵检测分类。该方法先通过 PCA 算法提取网络数据的关键成分，然后利用粒子群算法对 SVM 参数进行优化训练。算法同样提高了分类与攻击预测的准确性，但由于采用了 PCA、粒子群和 SVM 三类机器学习算法，导致复杂度较高，而且算法实际分类效果与 PCA 模型中专家经验有关。为提高入侵检测分类的准确性，文献[4] 引入双联 SVM 模型，尽管分类准确率提高，但是计算速度下降较多，并且双联 SVM 模型导致对于参数选择与优化更加敏感，很难实时自动选择合适的模型参数。

针对数据噪声问题，有学者引入模糊隶属度降低 SVM 模型对噪声点的敏感性[5]，但是该模型仅采用模糊处理和单级的 SVM 模型难以对网络安全数据中占比很低的攻击类数据进行高效多类别分类。为了保持分类的准确率、适应数据的噪声等问题，还有文献采用了其他机器学习模型与算法，主要包括：人工免疫模型[6]、K-均值聚类模型等。另一方面，入侵检测中的多类别模型研究也比较广泛，已有方法主要集中于直接采用多类 SVM 分类模型对原始数据集分类，如使用偏二叉树多类分类算法的 TWSVM 模型[7]、自适应阈值的多类 SVM 分类模型[8]、快速多类采样的 SVM 型等[9]。文献分析表明，支持向量机模型是入侵检测分类研究的主流模型。在基础 SVM 模型上，面向数据集合自身特征与分类的实际需求，设计合理的参数并融合其他手段可以有效进行入侵检测训练与分类。

相关研究表明，SVM 模型对核函数选择与参数调节非常敏感，这导致在多类分类算法中，SVM 的参数优化会随着类别的增加变得更加复杂。对于网络入侵检测来说，一方面训

收稿日期：2018-08-02；**修回日期**：2018-10-09

作者简介：汪生（1976-），男，安徽庐江人，高级工程师，博士，主要研究方向为网络空间安全；金志刚(1972-)，男，上海人，教授，博士，主要研究方向为网络信息安全等（zgjin@tju.edu.cn）.

练数据的类别分布不均匀，另一方面正常型样本数量远远超过攻击型样本数量。这导致直接使用原始数据训练单级 SVM 难以实现模型参数优化，产生的 SVM 分类器对于攻击型样本类别的分类准确率远低于正常样本分类的准确率。因此，为保证入侵检测多类判别的准确率，论文设计了模糊 SVM 模型的二分类与多值回归细分类模型相结合的新机制。首先训练模糊 SVM 对初始数据集进行二分类处理，然后将"攻击性数据"采用 DBscan 细分类模型进行多类自动聚类判别，得出攻击的具体类别。

1　基于 SVM 的粗细两级模型

1.1　模型整体设计

首先对数据包进行预处理，然后通过粗分类器进行判别，分类为"正常型数据"不需要进一步处理。对于"攻击数据"子集，使用细分类器分类为具体的攻击类型。

入侵检测分类模型分为两级：粗分类器对初始数据进行初筛选出感兴趣的数据；细分类器通过训练对初筛属于攻击的数据进行多类细分。如图 1 所示，预处理时将原始数据复制两份，第一份将正常型数据作为正样本，攻击型数据作为负样本，仅用于模糊 SVM 的粗分类；剔除正常型数据构造为第二训练集，经训练器训练后，得到适合于入侵检测的二值分类模型参数和回归模型多值分类的参数，然后统一将粗细分类训练参数传递到预测器，进行正式的分级多类判别。

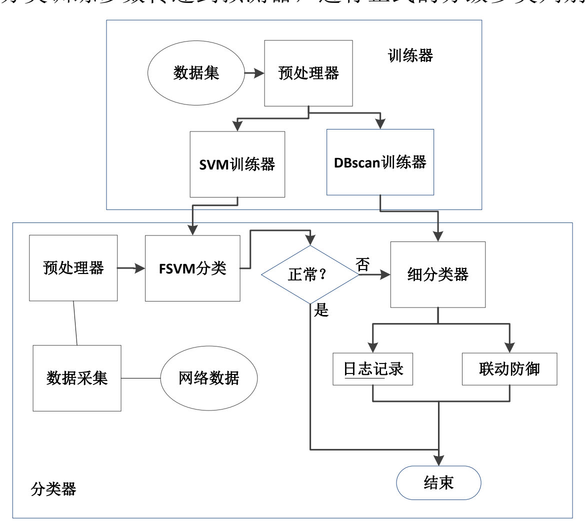


图 1　模型整体流程

Fig. 1　Flowchart of Classification

1.2　构建模糊分类模型

传统 SVM 模型容易受噪声点和孤立点影响导致分类不稳定，入侵检测的数据集往往噪声多，不能直接使用 SVM 模型。对数据点增加模糊隶属属性，通过引入模糊处理可以较好解决数据孤立点与噪声对分类的影响，用隶属度函数将 SVM 中的惩罚参数模糊化。通过优化训练 SVM 参数和调整隶属度获得针对入侵检测数据的高效粗分类模型。

对于特征向量为 v_i 标签为 l_i 的样本 i,增加函数 f_i 表示其隶属度，$0<f_i\le 1$；　将 SVM 的核函数表示为 $\phi(v)$，扩展后的模糊 SVM 模型最优分类等价为求解式(1) [10]。

$$\begin{cases} \min\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} f_i\xi_i \\ l_i[w\cdot\phi(v_i)+b]-1+\xi_i \ge 0 \end{cases} \quad i=1,2,\ldots,m \tag{1}$$

其中：ξ_i 是松弛因子，C 是固定常数，为了达到更好的优化结果可选择较小的 f_i 取值以降低对 ξ_i 优化结果的影响。采用拉格朗日松弛法将式(1)等价为求解式(2)的极值。

$$\begin{cases} \max W(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j l_i l_j K(v_i,v_j) \\ \sum_{i=1}^{m} l_i\alpha_i = 0 \quad 0\le\alpha_i\le f_i*C, \quad i=1,2,\ldots,m \end{cases} \tag{2}$$

对入侵检测数据的基于模糊 SVM 模型的二分问题转换为通过数据集的训练来优化模糊隶属度函数的问题。下面具体说明通过样本训练来优化隶属度函数，以获得粗分类模型参数的过程。

为了获得优化的函数 f_i，将训练集的标签分正常型和攻击型两个子集，对于正常子集 l_i 置为 1，攻击子集 l_i 置为-1。在计算样本 i 与样本子集中心的距离基础上，计算 f_i 。

正负样本的中心分别记作为 x_+ 和 x_- ，对于每个样本 i，计算样本与正样本中心和负样本中心的距离。以正样本为例，定义 $d_+ = \max_{\{v_i:l=1\}}\|v_+ - v_i\|$，同样可以计算负样本的 d-。进一步定义隶属度函数为

$$f_i = \begin{cases} \|v_+ - v_i\|/(d_+ + \delta) & l_i = 1 \\ \|v_- - v_i\|/(d_- + \delta) & l_i = -1 \end{cases} \tag{3}$$

对于全部样本，通过 LIBSVM 软件对隶属度函数 f_i 进行训练优化，将得到的模型参数应用到 1.3 小节的分类器中。

1.3　基于 DBSCAN 的攻击数据多类细分

DBSCAN(Density-based spatial clustering of application with noise)是一种基于密度的聚类算法[11]。DBSCAN 将数据点以密度为依据分为三类，判断依据是点的圆邻域内的数量是否超过 MinPts。邻域内点数超过 MinPts 数目的点称为**核心点**；在半径 Eps 内点的数量小于 MinPts，但是落在核心点的邻域内的称为**边界点**；其余被分类为**噪声点**。该算法的主要步骤如下：

a）建立新簇。对于未检查子集中的对象 p，如果 p 状态是未处理，则检查其邻域。判断邻域中的点数，如果数量不小于 MinPts，则建立新簇 C，并将其中的所有点加入候选集 N；

b）更新候选集。对候选集 N 中任一未处理的对象 q，检查其邻域，按照邻域点数判断。若至少包含 MinPts 个对象，则将这些对象加入 N；如果 q 不属于当前任何一个簇，则将 q 加入 C；

c）判断 N 是否为空。如果非空，则重复步骤 b)；

d）判断是否全部对象已经标记完。如果有剩余对象，返回步骤 c)。否则，处理结束。

总之，提出的新分类模型通过综合应用基于模糊 SVM 的二分类，再对被判别为估计的数据子集应用 DBSCAN 算法进行聚类，进一步将攻击数据分成不同类别的子类别，实现对包含孤立点和大量噪声，并且在整体数据集中占比较低的攻击数据的多类自动分类。

2　基于模糊 SVM 的二级入侵检测算法

为了对已经被粗分为攻击的数据进一步分类，采用一对多的多类分类模型，通过构造预期数量的分类进行训练。对于预测分类器，在预测某网络数据包属于何种攻击时，计算得到该网络数据包属于某类别的概率，然后概率最高所对应的类判断为该网络包所属类别。不同于仅直接使用 DBSCAN 算法分类，直接分类只应用密度信息分类数量受到数据来源和数据质量的影响[12]。论文设计的二级分类算法由于采用了模糊 SVM 进行处理，使得噪声点影响大大减少，可以通过 DBSCAN 算法自动生成聚类，获得稳定的类别数量。对于论文训练使用的数据集 KDDCup99 来说，可分为四类，因此只

需构造四个训练器。训练器基本流程如图 2 所示。

DBscan 算法的伪代码如下：

```
输入：数据对象集合 D，半径 Eps，密度阈值 MinPts
输出：聚类 C
DBSCAN（D, Eps, MinPts）
begin
init C=0; //初始化簇的个数为 0
for each unvisited point p in D
  mark p as visited; //将 p 标记为已访问
  N = getNeighbours (p, Eps);
  if sizeOf(N) < MinPts then
    mark p as Noise; //如果满足 sizeOf(N) < MinPts，则将 p 标记为噪声
  else
    C= next cluster; //建立新簇 C
    ExpandCluster (p, N, C, Eps, MinPts);
  end if
end for
end
```

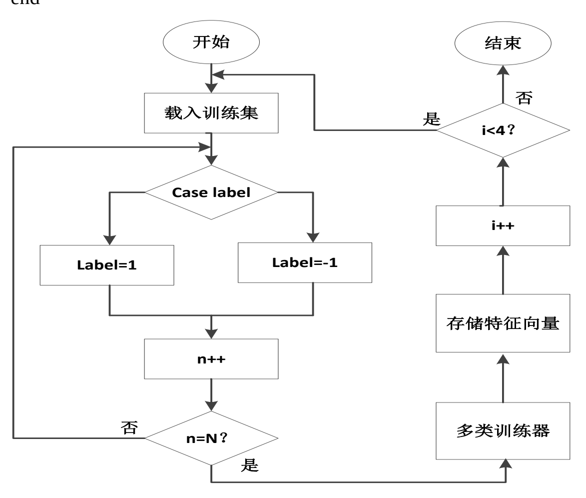


图 2　训练器基本流程

Fig. 2　Procedure of Trainer

其中关键步骤 ExpandCluster 的算法伪代码如下：

```
ExpandCluster(p, N, C, Eps, MinPts)
add p to cluster C; //首先将核心点加入 C
for each point p’ in N
  mark p' as visited;
  N’ = getNeighbours (p’, Eps); //对 N 邻域内的所有点进行半径检查
  if sizeOf(N’) >= MinPts then
    N = N+N’; //如果大于 MinPts，就扩展 N 的数目
  end if
  if p’ is not member of any cluster
    add p’ to cluster C; //将 p' 加入簇 C
  end if
end for
end ExpandCluster
```

3　实验验证

3.1　数据预处理

网络入侵检测系统直接收集的原始数据是网络中的二进制数据流，需要进行协议解析与格式转换才能用来分类。首先，使用 libpcap 库函数将二进制流解析为 IP 地址、端口和字符以及十六进制数值等字段。转换后的数据包含离散型字段和连续型字段，离散型字段又分为离散字符型和离散数字型。针对相关字段进行不同形式的预处理，转换为 SVM 模型可以接受并且噪声与误差少的加工数据。预处理流程包括三个步骤：离散字符型数据处理、数据标准与归一化和数据格式变换。

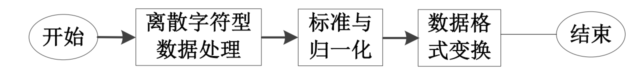


图 3　数据预处理流程

Fig. 3　The flow chart of data preprocessing

离散字符型数据处理将字符型数据转化为数值型数据以便于 DBSCAN 算法训练时计算距离等运算。数据标与归一化是原始数据的取值范围可能差距过大，导致的“大数吃小数”或者数据处理溢出以及权重不一致等问题。归一化处理消除度量单位对模型训练的影响，使训练结果更依赖于数据本身特征，从而提高聚类模型参数优化与分类预测的准确性。数据格式转换将经过数值化处理和归一化处理的数据进一步转换为 LIBSVM 支持的格式，以便进行模糊 SVM 模型的直接输入和训练。LIBSVM 格式广泛应用于常用的分类算法，一般采用{label　1:(value)$_1$　2:(value)$_2$　⋯⋯　i:(value)$_i$　⋯⋯　n:(value)$_n$}格式，其中 label 是类别标签，序号 i 是第 i 个字段的序号，(value)$_i$ 是第 i 个字段的数值。

下面采用 KDDCup99 数据集进行入侵检测系统的数据预处理。

3.1.1 字符型数据数值化

KDDCup99 数据集包含 42 个字段，其中 41 个是网络数据包特征属性字段，另外一个是该条数据记录的标签。为了避免字符数据不能直接计算距离以及过大过小数据直接应用影响平均值计算和距离计算效果等问题，将数据的字段按照表 1 的方法进行处理。

表 1　字符类数据数值化处理

Table 1　Quantizing of Chars

字段类别	TCP 连接基本特征(9 个)	标识特征	TCP 连接内容特征(13 个)	连续性数据
数值化	协议类型替换为整数	Flag 转换为十进制	Service 用端口号替换	保留原数值

例如，对 protocol_type 字段，其离散数值包括'TCP"UDP"ICMP 等，则'TCP’用 11 表示,'UDP'用 12 表示,'ICMP'用 20 表示。

3.1.2 归一化处理

对数值化处理过的数据进一步进行标准化与归一化处理，以同一特征属性为依据将训练集的数据具有同一特征的进行归一化。将带有标签的训练集中某条数据记为 x_{ij} ,i 表示训练集中数据的条数序号，j 表示特征的编号。通过计算特征 j 的平均值与方差来对数据进行标准化与归一化，计算公式如下：

$$\bar{x}_j = \frac{1}{n}\sum x_{ij} \tag{4}$$

$$S^2 = 1/n\sum (x_i - \bar{x})^2 \tag{5}$$

将数据用同特征的均值与方差来处理为标准数据 $x_{ij}^{'}$，

$$x_{ij}^{'} = \begin{cases} 0 & \bar{x}_j = 0 \text{或} S_j = 0 \\ \dfrac{x_{ij} - \bar{x}_j}{S_j} & \text{其他} \end{cases} \tag{6}$$

进一步进行归一化处理，数值归一化后的处理结果为

x_{ij}^{*}，归一化处理公式如下。

$$x_{ij}^{*}=\frac{x_{ij}^{'}-x_{\min}}{x_{\max}-x_{\min}} \tag{7}$$

3.1.3 数据格式转换

将经过数值化并且归一化的数据集转换为 LIBSVM 格式，以便进行 SVM 处理。论文使用的 KDDCup99 数据集，label 分为 5 种，即 NORMAL, DOS,R2L,U2R 和 PROBING。其中，NORMAL 代表正常的、没有攻击性的数据包；DOS 代表拒绝服务攻击；R2L 代表来自远程主机的未授权访问；U2R 代表未授权的本地超级用户特权访问；PROBING 代表端口被监视或扫描攻击。为了模型训练和预测分类需要，在数据格式转换时生成两组训练集：第一组将 NORMAL 数据的 Label 设置为-1，将其他四种数据的 Label 设置为 0；第二组将 NORMAL 数据剔除，其余数据保留。

经过上面的数值化处理、标准化与归一化处理以及数据格式转换就得到了实用数据集。

3.2　实验设计

为了训练和测试提出的攻击分类算法，搭建数据采集与安全分类实验环境（图 4），并通过 KDD CUP99 数据集驱动模拟攻击软件 IDS Informer 来产生不同种类的网络攻击数据包，客户端用来产生正常的网络业务数据。为了提高实验效率，通过对 KDD CUP99 集合均匀采样使用 50 万条数据的 20%来进行训练。也就是说利用正常与攻击行为共 10 万个训练样本进行模糊 SVM 参数优化和 DBSCAN 算法的聚类训练。粗细二级分类器训练完成后，对通过搭建的原型实验网络获取的真实业务截包，然后进行预测分类判别。表 2 是训练集的数据类别分布和搭建的实验网络获取的测试集的类别分布。

为了验证论文提出的算法的分类准确率和时间效率，首先仅使用优化训练的 DBSCAN 分类器与朴素贝叶斯分类器（NB）、支持向量机分类器（SVM）和随机森林分类器（RF）进行分类准确率对比。由图 5 可知，DBSCAN 分类器的准确率最高，可达到 85%，其他分类器的准确率均在 70%左右。

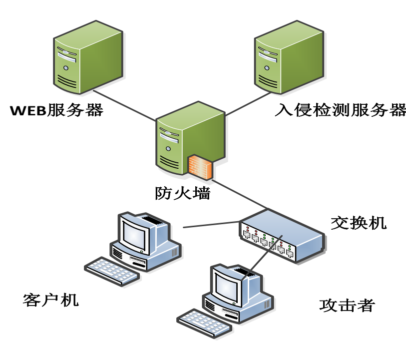


图 4　入侵检测实验环境

Fig. 4　Testing Environment of IDS

在训练时间方面，虽然朴素贝叶斯分类器的训练时间远小于其他三种分类器的训练时间，但是其分类准确率过低，并且对数据噪声过于敏感分类的类型数量不稳定。DBSCAN 分类器的训练耗时与随机森林分类器和支持向量机分类器的训练时间相近，具有稳定的分类数量和高分类准确率。

由于真实网络数据中异常的攻击数据占比很低，噪声明显，直接采用 DBSCAN 算法的分类效率偏低，而且训练时间过长。因此，进一步实验 2 中，采用表 2 所示分布的训练集和测试集，先使用模糊 SVM 二值分类器作为粗分类器，然后再用 DBSCAN 分类器对攻击型数据进行细分类，实验结果如表 3 所示。实验结果表明，模糊 SVM-DBSCAN 的联合两级分类器的准确率比单独使用一级 DBSCAN 分类器提高超过 10%，模型的训练时间只增长了约 30%。联合应用的粗细二级分类器更适合实际的网络场景进行实际业务数据的及时分类和联动入侵检测的应对措施。

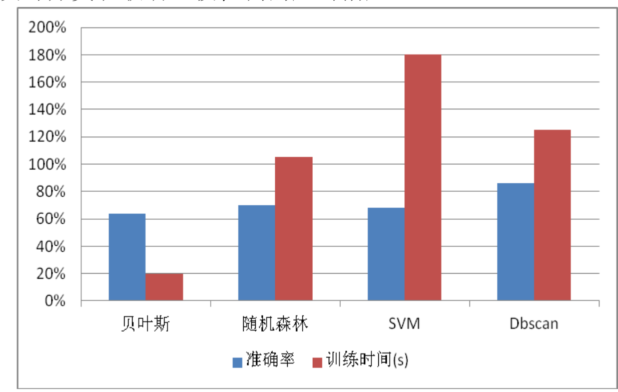


图 5　分类器对比实验

Fig. 5　Comparison of classsifiers

上述实验使用的是普通配置的电脑，如果将相关算法进行编译优化，并且运行到高配置的专用服务器并多线程执行，预计可以提高训练速度 6~8 倍，因此训练时间会下降到 0.2~0.25 s。进一步，如果将算法迁移到专用加速硬件可以至少提高训练速度一个数量级，也就是训练时间下降到 20~25 ms，该速度完全可以满足 IDS 系统的工作需要。

表 2　训练集与测试集数据类别

Table 2　Datasets of training and testing

类别	训练集	测试集
normal	20000	30000
DoS	4120	6276
probe	4000	6462
r2l	4496	5991
u2r	208	1500

表 3　分类器效果对比

Table 3　Results of classifiers

分类器	分类准确率	训练时间/s
模糊 SVM-DBscan 联合	96%	1.47
仅 DBscan	85%	1.25

4　结束语

在分析入侵检测与分类模型研究基础上，设计了面向多孤立点噪声数据的粗细联合分类模型，以模糊 SVM 分类为基础，对粗分为攻击的数据再利用 DBSCAN 模型进行聚类细分判别。实验结果表明，在训练样本分布不均匀、训练样本较小的情况下，DBSCAN 算法比其他多类分类模型的准确率高、训练耗时短。联合模糊 SVM 与 DBSCAN 的二级入侵检测算法在保持较快训练速度的前提下实现了高分类准确率，适用于正常型数据和攻击型数据分布不均匀的入侵检测。

参考文献：

[1] 中国互联网络信息中心，第 41 次《中国互联网络发展状况统计报告》[EB/OL]. [2018-03-15] http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/t20180305_70249. htm.

[2] 肖敏，韩继军，肖德宝，等. 基于聚类的入侵检测研究综述 [J]. 计算机应用, 2008,28 (s1): 34-38. (Xiao Min, Han Jijun, Xiao Debao, et al.Review of IDS research based on clustering[J]. Journal of Computer

Applications, 2008, 28(s1): 34-38)

[3] Kotpalliwar M V, Wajgi R. Classification of attacks using support vector machine (SVM) on KDDCUP'99 IDS database [C]//Proc of Internatiomal Conference on Communication Systems and Network Technologies. Piscataway, NJ: IEEE Press. 2015: 1-5.

[4] 王昊，华继学，范晓诗. 基于双联支持向量机的入侵检测技术 [J]. 山东大学学报: 工学版, 2013, 43(6): 53-56. (Wang Hao, Hua Jixue, Fan Xiaoshi, An Dual-SVM model based IDS scheme[J].Journal of Shandong University:Engineering Edition, 2013, 43(6): 53-56)

[5] Liu Jia, Fang Ning, Xie Yongjun, *et al.* Multi-scale feature-based fuzzy-support vector machine classification using radar range profiles [J]. IET Radar Sonar Navigation, 2015, 10(2): 370-378.

[6] 张玲，白中英，罗守山，等. 基于粗糙集和人工免疫的集成入侵检测模型[J].通信学报,2013,34(9): 166-176. (Zhang Ling, Bai Zhongying, Luo Shoushan,et al.Integrated IDS model based on rough set and artificial immunization[J].Journal of Communications, 2013, 34(9): 166-176)

[7] 谢娟英，张兵权，汪万紫. 基于双支持向量机的偏二叉树多类分类算法 [J]. 南京大学学报:自然科学版，2011,47(4): 354-363. (Xie Juanying, Zhang Bingquan, Wang Wanzi, Dual-SVM based partial binary tree classification algorithm[J]. Journal of Nanjing University:Natural Science Edition, 2011, 47(4): 354-363)

[8] Wan Shibiao, Mak Manwai, Kung Sunyuan. Adaptive thresholding for multi-label SVM classification with application to protein subcellular localization prediction [C]//Proc of IEEE International Conference on Acoustics. Piscataway, NJ: IEEE Press. 2013: 3547-3551.

[9] Chen Jingnan, Liu Chenlin. Fast multi-class sample reduction for speeding up support vector machines [C]//Proc of IEEE International Workshop on Machine Learning for Signal Processing Piscataway, NJ: IEEE Press,2011: 1-6.

[10] Li Shengtun, Chen C C. A regularized monotonic fuzzy support vector machine model for data mining with prior knowledge [J]. IEEE Trans on Fuzzy Systems, 2015, 23(5): 1713-1727.

[11] 王智罡，基于超像素的 SAR 图像海岸线检测算法 [D].大连:大连海事大学, 2017. (Wang Zhigang. Coastline detection algorithm of SAR image based on superpixel[D].Dalian:Dalian Maritime University, 2017)

[12] Zhang Minling, Wu Lei. Lift: multi-label learning with label-specific features [J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2015, 37(1): 107-120.